



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

농학석사학위논문

애기장대 종자 발달에 따른
유전체 프로파일링

Transcriptome profiling of seed developmental stages
in *Arabidopsis thaliana*

2019년 8월

서울대학교 대학원

농생명공학부 응용생명화학전공

김 제 민

A Dissertation for the Degree of Master of Science

**Transcriptome profiling of seed developmental
stages in *Arabidopsis thaliana***

August 2019

Jemin Kim

Applied Life Chemistry Major

Department of Agricultural Biotechnology

Seoul National University

애기장대 종자 발달에 따른
유전체 프로파일링

Transcriptome profiling of seed developmental stages
in *Arabidopsis thaliana*

지도교수 신 찬 석

이 논문을 농학석사학위논문으로 제출함
2019년 6월

서울대학교 대학원
농생명공학부 응용생명화학전공

김 제 민

김제민의 석사학위논문을 인준함
2019년 8월

위 원 장 김 민 균 (인)

부 위 원 장 오 기 봉 (인)

위 원 신 찬 석 (인)

ABSTRACT

Transcriptome profiling of seed developmental stages in *Arabidopsis thaliana*

Jemin Kim

Major in Applied Life Chemistry

Department of Agricultural Biotechnology

The Graduate School

Seoul National University

Seed dormancy and germination are the physiological phenomenon by a capacity between dormancy or germination to grow as the whole plants. These are determined by the narrow range of environments and hormones such as Absciscic acid (ABA) and Gibberellic acid (GA). Also, many studies have found the gene related to seed development such as Delay of germination1 (*DOG1*), in *Arabidopsis thaliana*. Recently, Non-coding RNA and long non-coding RNA were identified as the regulators in animals or plants.

Using strand-specific RNA sequencing data in this study, I performed that differential expression analysis and gene correlation analysis with *Arabidopsis*

thaliana seed samples in developmental stages. 6121 Protein-coding genes were up-regulated in After-ripening germination seeds, while 2636 protein-coding genes were down-regulated in Freshly harvested seeds. I classified protein-coding genes in developmental stages through expression pattern analysis and GO enrichment analysis. Most genes correlated to each marker gene. I identified 1199 novel lncRNAs in *Arabidopsis thaliana* seed. Of those, 417 lncRNAs were expressed with high abundance. 366 lncRNAs were expressed the opposite direction of protein-coding genes and correlated or anti-correlated with those. My study provides as guides to measure the criterion for dormancy and germination in seed developmental stages and as the fundamental data to identify the biological mechanism between protein-coding gene and long non-coding RNA in the cell.

Keyword: *Arabidopsis thaliana*, seed development, transcriptome profiling, protein-coding genes, lncRNA.

Student Number: 2017-21084

CONTENTS

ABSTRACT	i
CONTENTS.....	ii
LIST OF TABLES AND FIGURES	iv
LIST OF ABBREVIATIONS.....	v
INTRODUCTION.....	1
MATERIALS AND METHODS.....	5
RESULTS AND DISCUSSION.....	9
SUMMARY	16
FIGURES AND LEGENDS	17
TABLES.....	38
REFERENCE	42
SUPPLEMENTARY FIGURES AND LEGENDS	45
ABSTRACT IN KOREAN	48

LIST OF FIGURES AND TABLES

Figure 1.	Experimental scheme in this study
Figure 2.	Transcriptome profiling in <i>Arabidopsis thaliana</i> seeds
Figure 3.	Analysis of differential gene expression by time courses
Figure 4.	Analysis of differential gene expression among developmental stages
Figure 5.	Analysis of the expression patterns in developmental stages
Figure 6.	GO enrichment analysis in developmental stages
Figure 7.	Correlation analysis for novel candidates
Figure 8.	Profiling of lncRNA in <i>A. thaliana</i> seeds
Figure 9.	RT-qPCR analysis for validation of representative lncRNAs
Figure 10.	Reciprocal pattern analysis between protein-coding gene and lncRNA
Table 1.	List of primer sequences used in strand specific RNA-seq library preparation
Table 2.	Read distribution of strand specific RNA-seq library
Table 3.	List of primer sequences used in RT-qPCR validation
Table 4.	List of representative or putative genes in developmental stages

LIST OF ABBREVIATIONS

<i>A. thaliana</i>	<i>Arabidopsis thaliana</i>
DOG1	Delay of Germination1
FH	Freshly harvested
AR	After ripening
ARG	After ripening germination
ncRNA	Non-coding RNA
LncRNA	Long non-coding RNA
LincRNA	Long intergenic non-coding RNA
IncRNA	Intronic non-coding RNA
Cis-NAT	Cis-natural antisense transcript
FPKM _s	Fragments per kilobase of exon per million reads
GO	Gene Ontology
CPC2	Coding potential calculator2

INTRODUCTION

***Arabidopsis thaliana* and Seed development**

Arabidopsis thaliana (*A. thaliana*) is one of annual plants growing fast in numerous continents and is essential to experimental model in plant biology for studying plant development (Mitchell-Olds, 2001; Hoffmann, 2002).

The seed development is determined by the relative difference of a capacity between dormancy and germination in the narrow range of environmental changes. The seed consists of the following compartments: seed coat preventing the embryo from mechanical wound or microbes, endosperm persisting to the mature seed stage as a storage tissue and embryo consisting of cotyledon, epicotyl, plumule, hypocotyl, radicle. Seed requires coordinate processes such as differentiation, development and maturation to grow it properly. The seed developmental stages are classified by the following three stages according to the maturation period: Freshly harvested (FH), After-ripening (AR), After ripening germination (ARG). FH are green seed of the premature status in the process of maturation and has a low capacity for germination in the effect of strong dormancy. AR is yellow seed of the matured status in the dry seed, which is regulated by a capacity for dormancy or germination (Bewley, 1994). ARG is yellow-brown seeds of germination-possible status by seed imbibition at a low temperature called stratification.

Arabidopsis thaliana *DELAY OF GERMINATION 1* (*AtDOG1*) is a core protein to regulate seed dormancy for adapting uncertain environments (Alonso-Blanco et al 2003). *AtDOG1* mutants were completely non-dormant status (Bentsink et al., 2006, 2010). *DOG1* regulated a conserved physiological coat-dormancy mechanism

in the Brassicaceae *L. sativum* and *A. thaliana* (Graeber et al., 2014). Also, seed development requires hormonal regulation which is a conserved mechanism in the seed. Absciscic acid (ABA) and gibberellin acid (GA), which is well-known hormones typically in seed development, maintain in a capacity for dormancy and germination. those were catalyzed by nine-cis-epoxycarotenoid dioxygenase (NCED) or gibberellin acid 3-oxidase (GA3OX).

This study, then, has focused on which whether genes in *A. thaliana* seed are up-regulated or down-regulated in developmental stages as comparing gene expression levels.

Non-coding RNA

Non-coding RNA (ncRNA), which is not translated into proteins, was thought to junk RNA in the past, but many studies have identified the diverse roles of those from a variety of species (Morris & Mattick, 2014). ncRNA is classified by their molecular mechanism in cellular biology (Hombach & Kretz, 2016): microRNA (miRNA), small interfering RNA (siRNA), Piwi-interacting RNA (piRNA), long non-coding RNA (lncRNA), etc. ncRNA is necessary for a regulatory role to fundamental biological processes such as cell growth, immune reaction, response to a variety of stresses (Zhao et al., 2007; van Rooji et al., 2007). Further, it is associated with a variety of diseases such as cancer or Alzheimer's disease (Gacal et al., 2002; Wang et al., 2008). Large scale analysis of ncRNA function in these studies would provide for understanding and defining their roles in the cell.

Long non-coding RNA

Long non-coding RNA (lncRNA) which is transcripts more than 200 nucleotides in length have been identified as important regulators of diverse biological processes (Mercer et al., 2009; Wilusz & Sunwoo, 2009). The structure of lncRNA is similar to that of messenger RNA (mRNA), which undergoes post-transcriptional modification such as 5' capping, RNA splicing, and 3' polyadenylation, but is not translated to protein. LncRNA is mainly classified into four types (Briggs, James et al., 2015; Iyer et al., 2015): natural antisense transcript lncRNA (NAT-lncRNA), overlapping lncRNA (OT-lncRNA), intergenic non-coding RNA (lincRNA), and intronic non-coding RNA (IncRNA). NAT-lncRNA is transcribed to the opposite direction of their protein-coding genes. OT-lncRNA is overlapped partially or completely to the sense direction of their protein-coding genes. LincRNA is generated from intergenic regions, while IncRNA is generated from intronic regions.

In recent years, lncRNA was found in a variety of species (Cheng et al., 2005; Katayama et al., 2005; Henz et al., 2007; Wang et al., 2014). Indeed, novel lncRNA was identified genome-wide sequencing analysis using RNA-seq data in human (Ma et al., 2018). Many studies found that it was associated with various human cancers (Ning et al., 2015; Hajjari et al., 2015). In plants, novel lncRNA is identified in *A. thaliana* (Wang et al., 2005, 2014), maize (Li et al., 2014; Wang et al., 2015), rice (Lu et al., 2012; Zhang et al., 2014; Wang et al., 2015) through transcriptomic analysis. For example, COOLAIR is transcribed to spliced alternatively and polyadenylated transcript from the FLOWERING LOCUS C (FLC) locus at an early stage of cold exposure (Heo & Sung et al., 2011; Kim & Sung et al., 2017), *DROUGHT INDUCED lncRNA* (DRIR) is significantly activated by drought and salt stress as well as by abscisic acid (Qin et al., 2017).

Purpose of study

Many studies have identified the biological mechanism of protein-coding genes in *A. thaliana* seed. However, these mainly analyzed for a particular gene expressed in FH or AR seed. In this study, I supplemented ARG seed to obtain protein-coding genes expressed in developmental stages, accurately. I performed transcriptomic analysis on RNA-seq data and compared the expression levels of protein-coding genes in developmental stages to select putative genes for measuring a capacity for dormancy or germination. Also, I identified novel lncRNA related to protein-coding gene. On the basis of the results, I would provide fundamental data to identify the biological mechanism in the cell by suggesting the interaction between protein-coding gene and lncRNA in developmental stages in *A. thaliana* seed.

MATERIALS AND METHODS

Plant materials and growth condition

A. thaliana seeds in this study were used in the Columbia (Col-0) background. All plants were grown in the chamber with a cycle of 16 h light / 22 °C and 8 h dark / 18 °C. FH seeds were harvested from 4-week-old seedling. For obtaining AR seeds, the seeds harvested from 5-week-old seedling were dried for 2 weeks. For inducing seed germination from AR seeds, first, the seeds were sterilized with a solution containing 0.08 % TritonX-100, 70 % EtOH for 30 min and were washed with 100% EtOH. Second, the seeds dried on filter paper for 20 min and were sealed with aluminum foil and stored at 4 °C for 3 days to proceed stratification. Finally, the imbibed seeds were incubated for 2 h in the chamber with 100 μ mol of intensity of light (22 °C) and for a day in the chamber with 24 h dark cycle (22 °C).

RNA isolation

A. thaliana seeds were ground with cold mortar and pestle. each 50 μ g of ground seeds was washed with EB buffer containing 100 mM Tris-HCl (pH 9.5), 150 mM NaCl, 1.0 % Sarkosyl, 5 mM DTT. The pellets were collected by centrifugation at 15,000 \times g for 10 min at 4 °C and were purified using phenol-chloroform extraction. Total RNA was extracted with TRIzol (Invitrogen) and treated with DNase I (Takara) to eliminate genomic DNA. The quantity of total RNA was calculated with Nanodrop 2000 spectrophotometer and the purity was confirmed using non-denaturing agarose gel electrophoresis.

Library preparation for strand-specific RNA sequencing

Poly-adenylated RNA was isolated from Total RNA purified with Oligo(dT) beads (Invitrogen) to prepare cDNA library. All libraries were constructed by using user-supplied strand-specific library protocol applying dUTP method (Zhong et al., 2011).

Transcriptome reconstruction

For reconstructing *A. thaliana* Transcriptome from ribo-minus RNA sequencing data, raw reads (quality score ≥ 30) were aligned to *Arabidops thaliana* reference genome (TAIR10) databases (Lamesch et al., 2012) using HISAT2 (Pertea et al., 2016). Mapped paired-reads were assembled into transcripts by Stringtie (Pertea et al., 2015). Merged transcripts were confirmed how it is recovered to reference genome annotation. Annotated transcripts were counted with the parameter (FPKM, Fragments per kilobase of exon per million reads) by Stringtie.

LncRNA identification

To identify novel lncRNAs systemically in the *Arabidopsis thaliana* seed, we referred to online protocol (Lu et al., 2014). First, to discover lncRNAs, we removed the same strands, homologies, and non-coding RNAs overlapped to Araport11 annotation (Cheng et al., 2017). Putative lncRNAs were discovered by homology of Araport11 lncRNA. Second, for analyzing the structure of putative lncRNAs, we compared putative lncRNAs with Rfam.cm data of Rfam ver.13 (Kalvari et al., 2017). Transcripts similar to already discovered lncRNAs were filtered using infernal version 1.1.2 (Nawrocki and Eddy, 2013). Third, for removing transcripts possible to translate into proteins, we calculated coding potential for those using coding

potential calculator2 (Kang et al. 2017). Next, we removed Transcripts with short length (length < 200-nt) or low abundance (FPKM < 2 based on raw reads). Finally, Putative lncRNAs were obtained by mapping transcripts in *A. thaliana* genome.

RT-qPCR for validation of lncRNA

First-strand transcript was synthesized using 2 µg of total RNA with oligo(dT) primer and superscript III (Invitrogen). To confirm expression levels of putative lncRNAs expressed significantly in developmental stages, RT-qPCR was performed using LightCycler® 480 SYBR Green I Master (Roche) with gene-specific primer sets. expression levels of those were normalized to the expression level of UBQ11. Primer sequences are attached in Table 3.

Gene Ontology (GO) analysis

To classify up-regulated genes in developmental stage by biological process, GO enrichment analysis was performed using genes with the high abundance (FPKM \geq 2). For avoiding high false discovery rate (FDR) in multiple testing, q-value (Storey, 2002) was also estimated for FDR control (FDR < 0.05). All results were calculated systemically based on gene classification method by The clusterProfiler in R (Guangchuang Yu, 2012).

Gene correlation analysis

Gene correlation analysis was performed using Pearson correlation coefficient between protein-coding genes or between protein-coding gene and lncRNA. Protein-

coding genes with strong positive coefficient ($R \geq 0.85$) were compared to marker gene.

RESULTS AND DISCUSSION

Profiling of transcript in *A. thaliana* seeds

To confirm whether the gene expression on RNA-seq data is accurate in developmental stage before I perform profiling of transcriptome in *A. thaliana* seed, I confirmed the expression levels of marker genes related to dormancy and germination using FPKM value normalized by reference genome. *NCED* catalyzing biosynthesis of ABA were relatively up-regulated in FH compared with AR and ARG (Supplementary Figure 1A, 1B, 1C and 1D). Cytochrome P450, family 707, subfamily A, polypeptide 2 (*CYP707A2*) encoding abscisic acid 8'-hydroxylases was up-regulated in AR compared with FH and ARG (Supplementary Figure 1F). *GA3OX1* and *GA3OX2* catalyzing biosynthesis of GA were up-regulated in ARG (Supplementary Figure 1G and 1H). Finally, *DOG1* regulating seed dormancy was up-regulated significantly in FH (Supplementary Figure 1E). This indicated that marker genes in developmental stages were expressed accurately on RNA-seq data.

To classify annotated genes in each sample before I compare differential expression levels of protein-coding genes among 3 samples, transcript profiling was performed systemically. Figure 1A showed that 23,379 transcripts were obtained by RNA-seq assembly. The number of annotated transcript was 14,351 loci and 3,317 isoforms, the number of alternative spliced (AS) transcript was 3,444 loci and 857 isoforms, antisense transcript was 485 loci and 43 isoforms, the number of other transcript was 768 loci and 114 isoforms (Figure 2A). Most transcripts retained one exon and AS transcripts retained three exons significantly. (Figure 2B). Annotated and AS transcripts had a peak in 1800 nt (nucleotide), while antisense and other

transcripts had a peak in 500 nt (Figure 2C). Interestingly, non-annotated AS transcripts had a similar pattern for length and the number of exons of annotated transcripts. Also, antisense and other transcripts were shorter and smaller than annotated transcripts (Figure 2B and 2C). This indicated that AS transcripts are possibility of the isoforms derived from the same region of annotated transcripts. Also, antisense and other transcripts are possible to lncRNA. Annotated transcripts were expressed significantly in total transcripts, while antisense transcripts were expressed lower than other transcripts (figure 1D). By these results, I performed differential expression analysis using annotated transcripts in developmental stages.

Gene expression levels were up-regulated significantly in ARG

To compare the expression levels of annotated transcripts by time course, gene expression levels between each 2 developmental stages were calculated relatively using annotated transcripts with the abundance (FPKM > 1) and in the range of 3-fold changes. The results showed that there are 1442 up-regulated genes in FH and 4637 up-regulated genes in AR (Figure 3B). The distribution of genes was up-regulated in AR than in FH, relatively. (Figure 3A and 3B). There are 1914 up-regulated genes in AR and 4650 up-regulated genes in ARG (Figure 3B). genes were distributed significantly in ARG, relatively (Figure 3A and 3B). Also, there are 1030 up-regulated genes in FH and 6714 up-regulated genes in ARG (Figure 3B). The distribution of genes was significantly high in ARG (Figure 3A and 3B). This indicated that the expression levels of annotated genes were relatively up-regulated as seed growth

Next, the number of expressed genes were compared relatively among 3 developmental stages. Figure 4A showed that 6121 genes were up-regulated in ARG,

while 596 genes in FH were up-regulated. As expected, gene expression levels in ARG were up-regulated rather than in FH and AR. While, the number of down-regulated genes was 2636 genes in FH and 896 genes in ARG (Figure 4B). I found that gene expression levels were up-regulated or down-regulated relatively in narrow range of expression levels (Figure 4C).

A variety of expression pattern in developmental stages

To find putative genes possible to measure a capacity for dormancy or germination, the patterns of gene expression were analyzed using annotated genes with the high abundance ($\text{FPKM} > 3$) in each sample. The result showed that there were 8 patterns for gene expression (Figure 5). The number of annotated genes showing high expression levels in each developmental stage was 172 genes in FH, 756 genes in AR, and 3564 genes in ARG (Figure 5A, 5B and 5C). The number of co-expressed genes were 142 genes in FH and AR, 148 genes in FH-ARG, 906 genes in AR-ARG (Figure 5D, 5E and 5F). The number of down-regulated or up-regulated genes in developmental stages were 47 and 379 (Figure 5G and 5H). To classify biological process of genes from this result, I selected annotated genes which are up-regulated in each developmental stage.

Different biological process in developmental stages

To classify biological process of genes in developmental stages, GO enrichment analysis were performed using annotated genes with high abundance ($\log_2(\text{FPKM} + 1) > 2$) in each sample. The results showed that biological processes such as responses to wounding, chitin, and organonitrogen compound were grouped in FH (Figure 6).

This indicated that the death of a numerous cell in FH seeds were caused by the detrimental factor such as bacteria or virus during a process of seed maturation, those were expressed to protect the pre-mature seeds from those microbes during seed dormancy (Hadrami et al., 2010). Also, growth-related genes such as ethylene were included in the top list of GO analysis in FH (Figure 6), indicating that those were expressed highly to mature the seeds although the FH seeds were regulated by ABA and DOG1 to sustain the quiescence of seed growth. The genes expressed significantly in AR were mainly included in the biological process responding to heat, toxic substance, oxidative stress or water (Figure 6). This indicated that the seeds during the dry process were removed moisture considerably or were exposed to external environments directly. Therefore, those genes related to response to these stresses were expressed to adapt to environments (Murthy et al., 2002). Most genes in ARG were included in the groups catalyzing biosynthetic processes such as carbohydrate catabolic process, ribonucleotide metabolic process (Figure 6). This indicated that essential chemical elements were generated by using the nutrients stored during seed dormancy or the process of maturation. With this result, I found that there are clear differences for the biological processes of annotated genes which were up-regulated in developmental stages, assuming that up-regulated genes are significantly difference to adapt to encountered environments in developmental stages. Therefore, I selected annotated genes in accordance with the following three conditions: First, a group ranked in top list from the results of GO analysis. Second, a group related to marker gene. Third, a group responded to growth hormone.

Up-regulated genes in developmental stages were correlated to marker gene

To identify whether selected genes correlate to marker genes, correlation analysis

were performed using Pearson correlation coefficient. This result showed that the coefficient in FH was 0.97, indicating that there is positively strong correlation to *MAPKKK14* (Figure 7A). The coefficient in AR was 0.94, indicating that there is correlation to *CCD7* (Figure 7A). the coefficient was 0.97 in ARG, indicating that there is positive correlation significantly to *GA3OX2* (Figure 7A).

I compared its expression levels to the expression levels of marker gene with this result. Of those, 8 genes in each developmental stage were selected, which is similar or higher expressed than marker gene. protein-coding genes such as *ERF4* or *RAP2.4* (ethylene-activated signaling pathway), *RNS1* or *WIP3* (response to wounding), and *WRKY* family genes were more up-regulated significantly than the expression level of marker gene in FH (Figure B and table 4A). protein-coding genes such as *GOLS2* (response to water deprivation), *SKP2B* or *NAC089* (negative regulation of development), *SKIPI* (protein ubiquitination) were up-regulated significantly in AR (Figure B and table 4B). growth-related genes such as *SPR1* or *LNG1* (cell growth), *CYP705A5* (root development), and *DWF1* (brassinosteroid biosynthetic process) were up-regulated significantly in ARG (Figure B and table 4C).

Identification of lncRNA in *A. thaliana* seed

I performed the following steps to identify novel lncRNAs in *A. thaliana* seed (Figure 8A). the result showed that 1420 loci of 1499 transcripts is no possible to encode protein-coding gene (Figure 8B). 417 loci of total 1420 putative lncRNAs were with the abundance (FPKM>0.5). 366 loci of 417 lncRNAs were expressed in antisense direction of protein-coding gene, 34 loci were intergenic RNA, 17 loci were intronic RNA. Also, there are 259 protein-coding genes with the abundance (FPKM > 1) overlapping to putative lncRNA, 33 intergenic lncRNAs were adjacent

to protein-coding gene within 50 kb with abundance (FPKM >1), and 14 intronic lncRNAs were adjacent to protein-coding gene (figure not shown). Antisense and intergenic lncRNAs had a peak in 500 bp, while intronic RNA had a peak in 250 bp (figure 8C). All lncRNAs were distributed from chromosome 1 to 5 (Figure 8D). Interestingly, I found that All lncRNAs were expressed in mitochondrial and chloroplast chromosome, suggesting that lncRNAs have a diverse function in biological processes.

RT-qPCR validation of putative lncRNAs

To validate whether lncRNAs on RNA-seq was expressed in practice, RT-qPCR and regression analysis was performed using putative cis-NATs overlapping to representative protein-coding genes related to development or stress response including *asDOG1* (Fedak et al., 2016). The results showed that the $\log_2(\text{FPKM}+1)$ values of *MSTRG.5971* appeared 10-fold decrease in FH and AR, while the relative expression levels were calculated similarly in those (Figure 9A and 9B). This indicated that ΔCT value of those calculated 40 that means those were not expressed (figure not shown). But regression coefficient had value of 0.993 which means those in RNA-seq were expressed similarly in developmental stages (Figure 9C). The relative expression levels of *MSTRG.22786*, *MSTRG.9541* and *asDOG1* appeared the similar aspects compared to $\log_2(\text{FPKM}+1)$ values of those on RNA-seq (Figure 9A and B). Also, regression coefficients for value of RNA-seq and RT-qPCR showed strong positive sign of those (Figure 9C). This indicated that validation of RT-qPCR provide strong evidence for lncRNAs expression on RNA-seq data.

LncRNAs were correlated or anti-correlated to protein-coding gene

Recently, the study reported that *asDOG1* is expressed in *Arabidopsis* seed and suppress seed dormancy against *AtDOG1* (Fedak et al., 2016). To investigate the interaction between lncRNA and protein-coding gene which used in genetic marker or expressed significantly on RNA-seq, I calculated Pearson correlation coefficient using those with the abundance (FPKM > 1). The results showed that *DIP2*, *IRX14* and *SKIP1* had high correlation coefficients to their lncRNA (R=0.97, 1 and 0.99), assuming that those aid or enhance the function of protein-coding gene (figure 10A, 10B and 10C). *CYP707A2* was correlated slightly (R=0.49) to *MSTRG.8135* acting in cis-configuration (figure 10D). Interestingly, the interaction was anti-correlated between FH and AR, assuming that *MSTRG.8135* enhances the expression of *CYP707A2* during process of seed maturation from FH to AR. *NCED4* and *MSTRG.16329* was anti-correlated in ARG (R=-0.69), assuming that *MSTRG.16329* regulates the expression of *NCED4*, which suppresses seed dormancy to proceed seed germination in ARG (figure 10E). *GSDLI* catalyzing lipid biosynthesis was anti-correlated to *MSTRG.2753* (R=-0.92), assuming that *MSTRG.2753* inhibits *GSDLI* to regulate cell growth in developmental stages (Figure 10F). This result indicated that lncRNAs were correlated or anti-correlated to protein-coding genes. Also, I predicted lncRNAs conduct a specific function in *A. thaliana* seeds. The predicted model for the mechanism of protein-coding gene and lncRNA in this study would provide as fundamental data to measure a capacity for seed dormancy or germination.

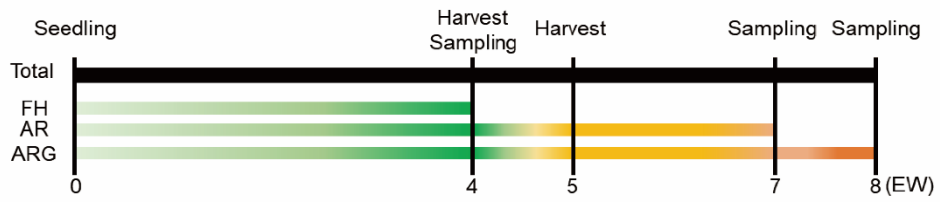
SUMMARY

Given with the raw data on RNA-seq, I performed large-scale transcriptomic profiling in developmental stages in *Arabidopsis thaliana* seeds. The profiling and the differential expression analysis of transcripts were performed using strand-specific RNA-seq data, relatively. Number of annotated genes were calculated up-regulated or down-regulated from developmental stages. Annotated genes were up-regulated significantly in ARG compared with FH and AR. In otherwise, there were more down-regulated genes in FH than in AR and ARG. Next, annotate genes were selected by performing GO enrichment analysis to classify biological process of those. Putative candidates to calculate capacity to dormancy or germination were confirmed by comparing to marker genes in each pattern.

Genome-wide identification of putative lncRNAs were performed using CPC with strand-specific RNA-seq data. Total 1420 putative lncRNAs were identified, of those, lncRNAs overlapped or adjacent to annotated genes were 417 antisense lncRNA, 34 lincRNA, and 17 incRNA. By performing RT-qPCR validation analysis, the expression levels of lncRNA were confirmed. Finally, by calculating correlation coefficient with annotated genes and lncRNAs to identify the interaction between those, I would provide fundamental data for their interaction in developmental stages.

FIGURES AND LEGENDS

(A)



(B)

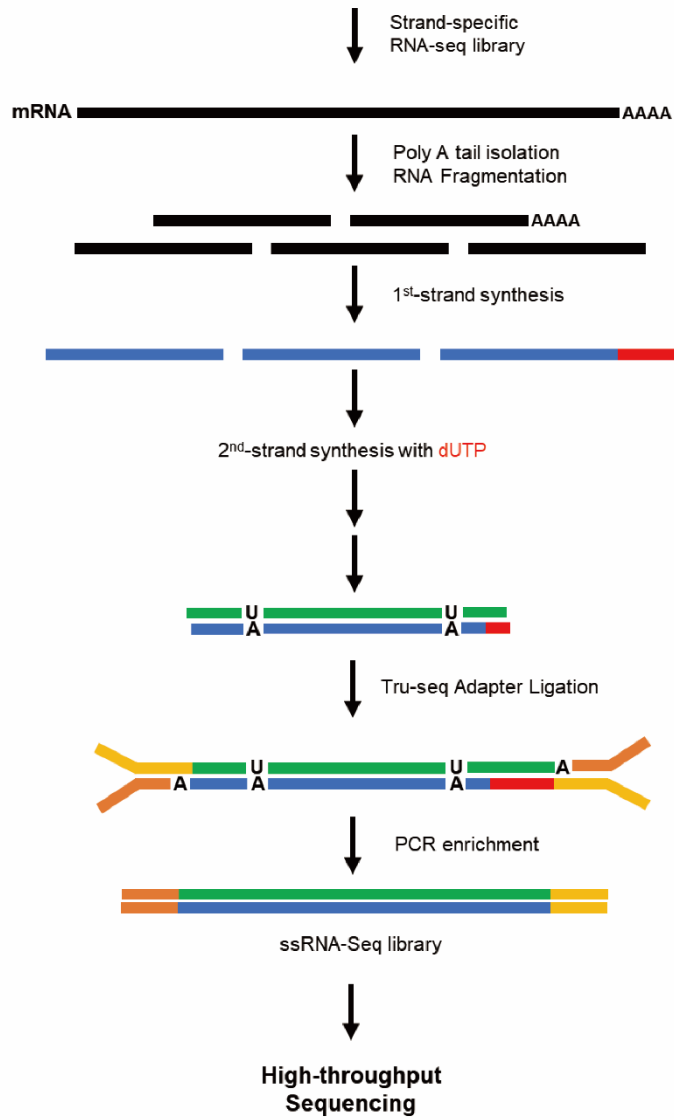


Figure 1. Experimental scheme in this study. (A) Scheme of the seed harvest in developmental stages. (B) Scheme of Strand-specific RNA-seq library preparation.

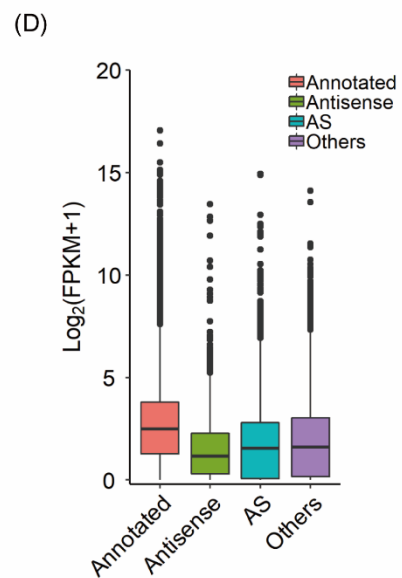
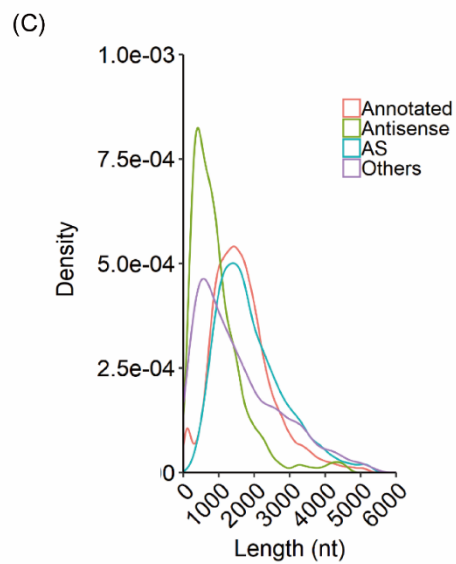
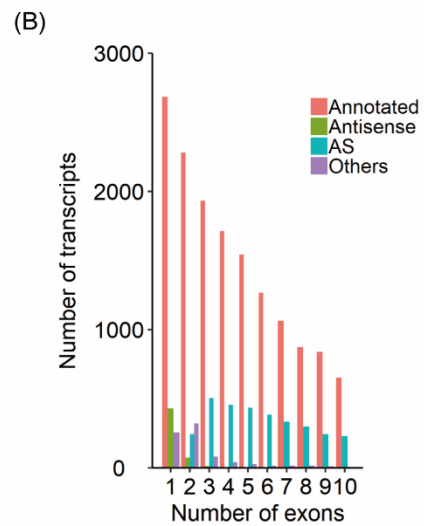
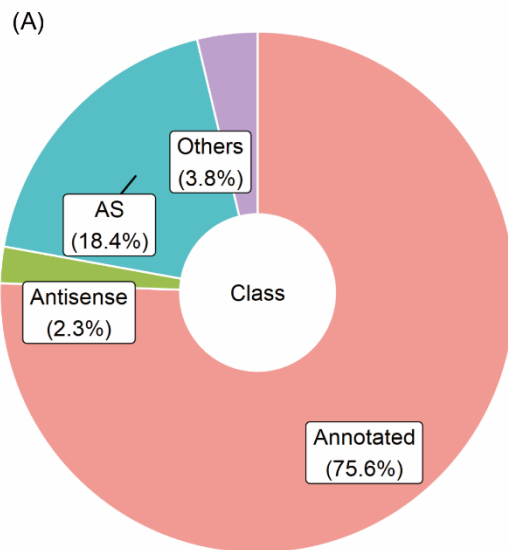


Figure 2. Transcriptome profiling in *Arabidopsis thaliana* seeds. The subgroups of transcripts were classified as annotated, antisense, alternative splicing (AS) and Others. (A) Proportion of transcripts. (B) Number of exons of transcripts. (C) Length distribution of transcripts (D) distribution of gene expression levels of transcripts.

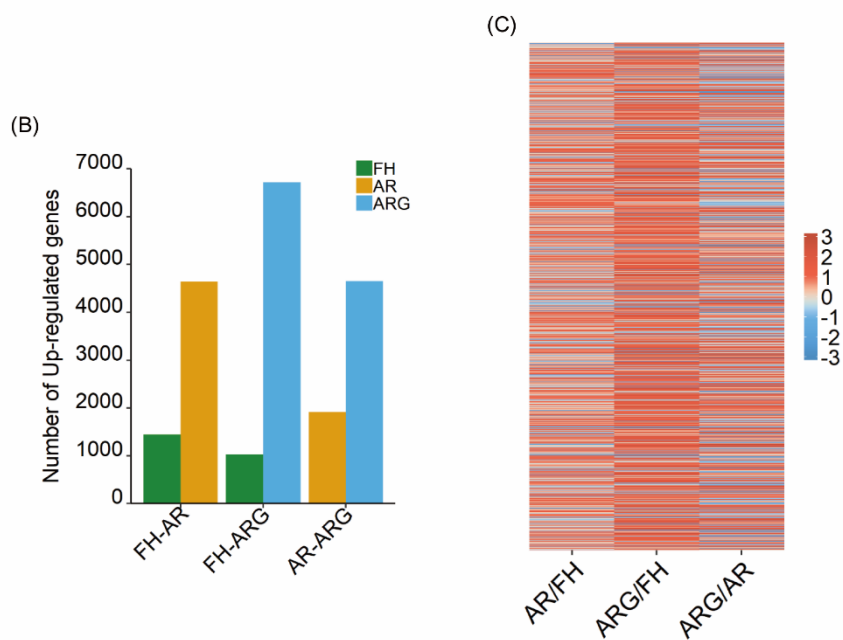
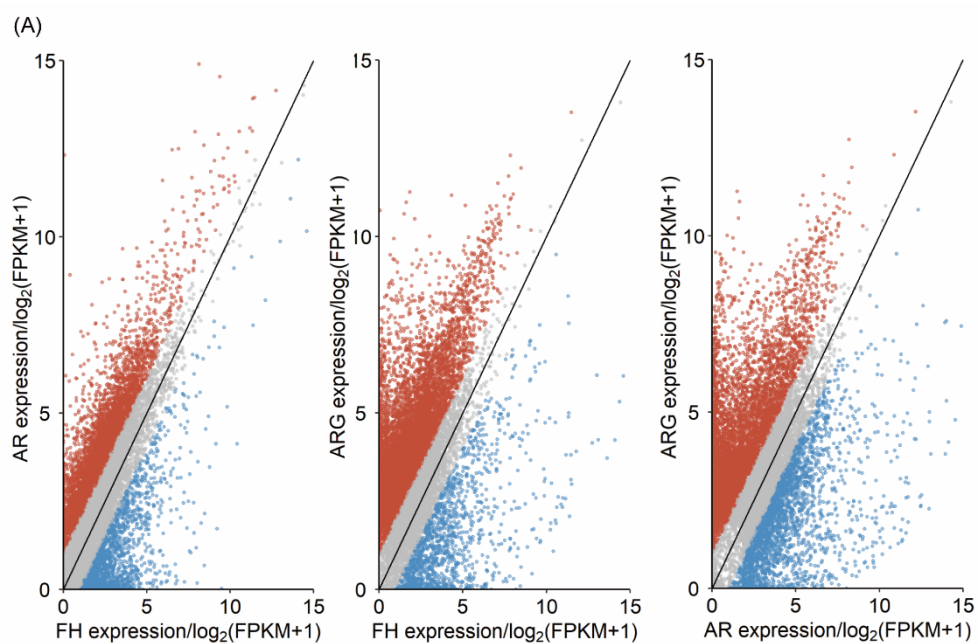
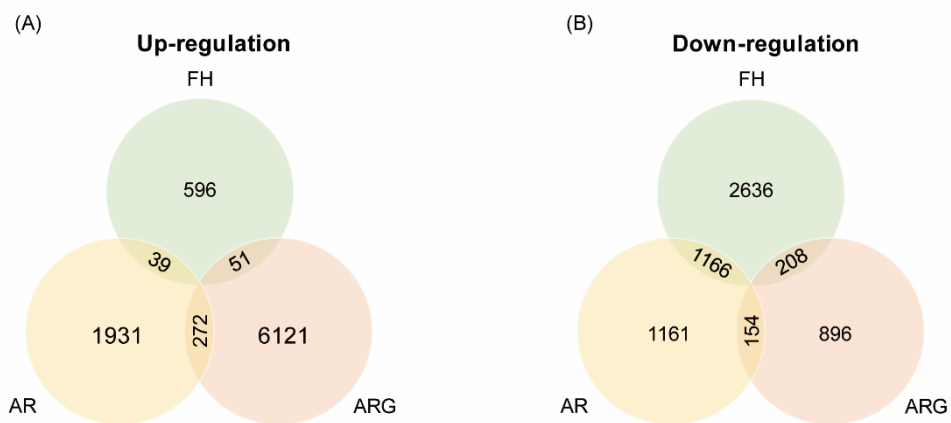


Figure 3. Analysis of differential gene expression by time courses. (A) Comparison of log-fold change in developmental stages. (B) Number of ≥ 2 -fold-upregulated genes in developmental stages. (C) heatmap representative of



(C)

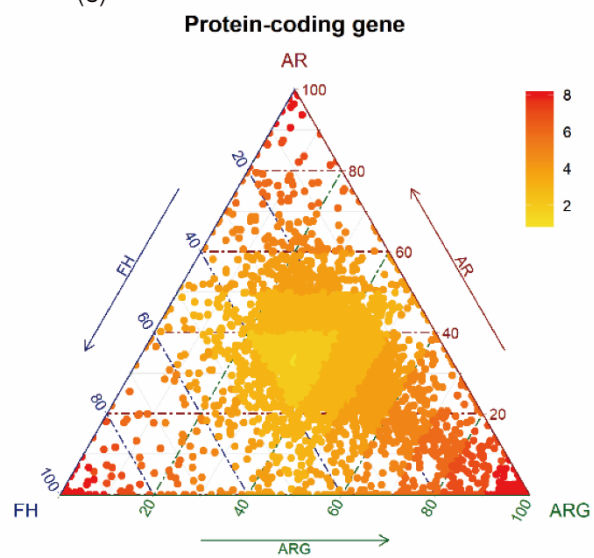


Figure 4. Analysis of differential gene expression among developmental stages. Number of ≥ 2 -fold (A) up-regulated and (B) down-regulated genes among 3 developmental stages. (C) distribution of gene expression among 3 developmental stages.

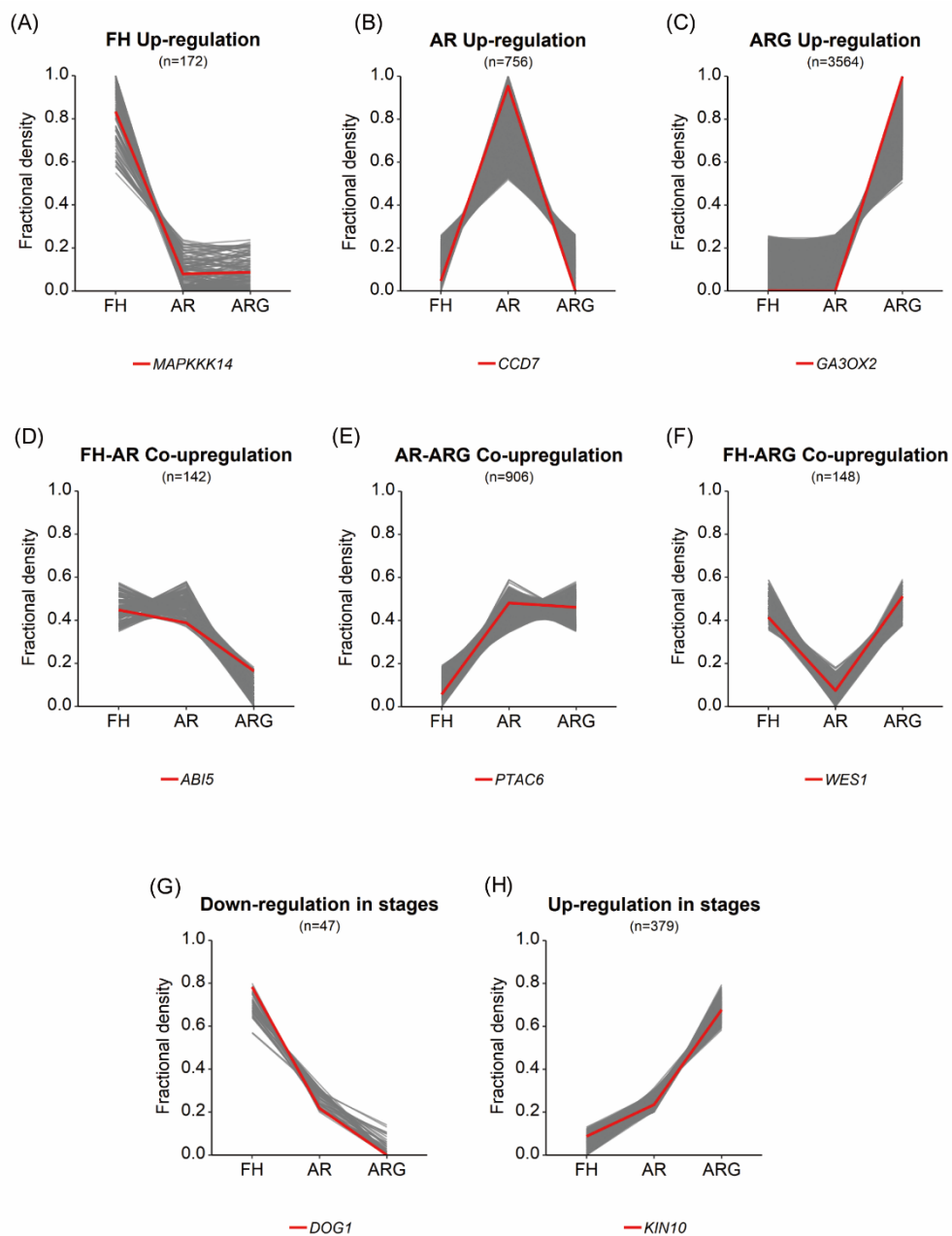


Figure 5. Analysis of the expression patterns in developmental stages. Genes in all patterns are with expression levels (FPKM > 3) (A)(B)(C) Number of upregulated genes in each stage. (D)(E)(F) Number of co-upregulated genes in two stages. Number of (G) down-regulated or (H) up-regulated genes in stages

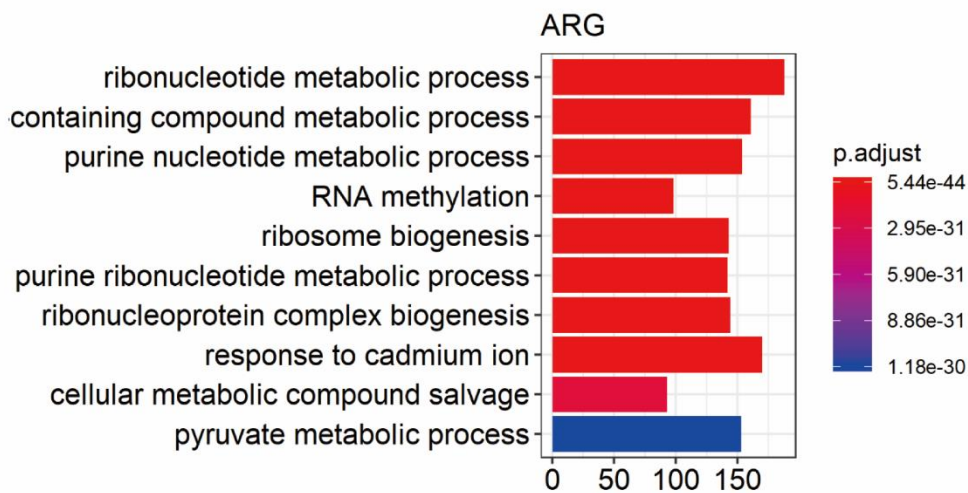
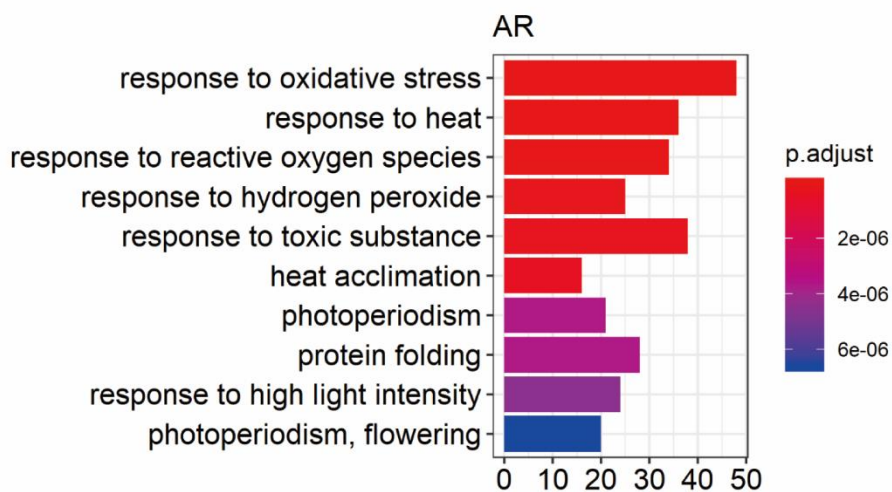
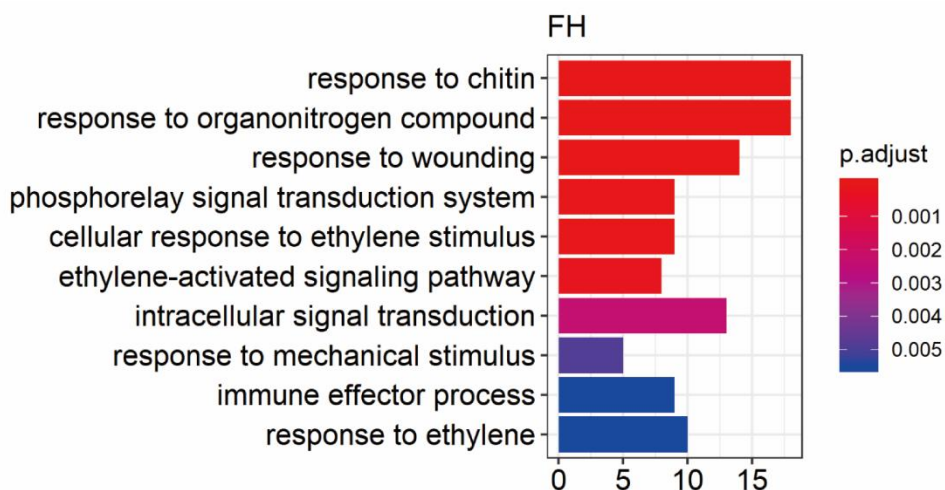


Figure 6. GO enrichment analysis in developmental stages. Biological process of annotated genes in developmental stages. the distribution of GO terms was exhibited statistical significant differences (Fisher Exact Test, $P < 0.05$, FDR < 0.05 , 2-fold cut-off).

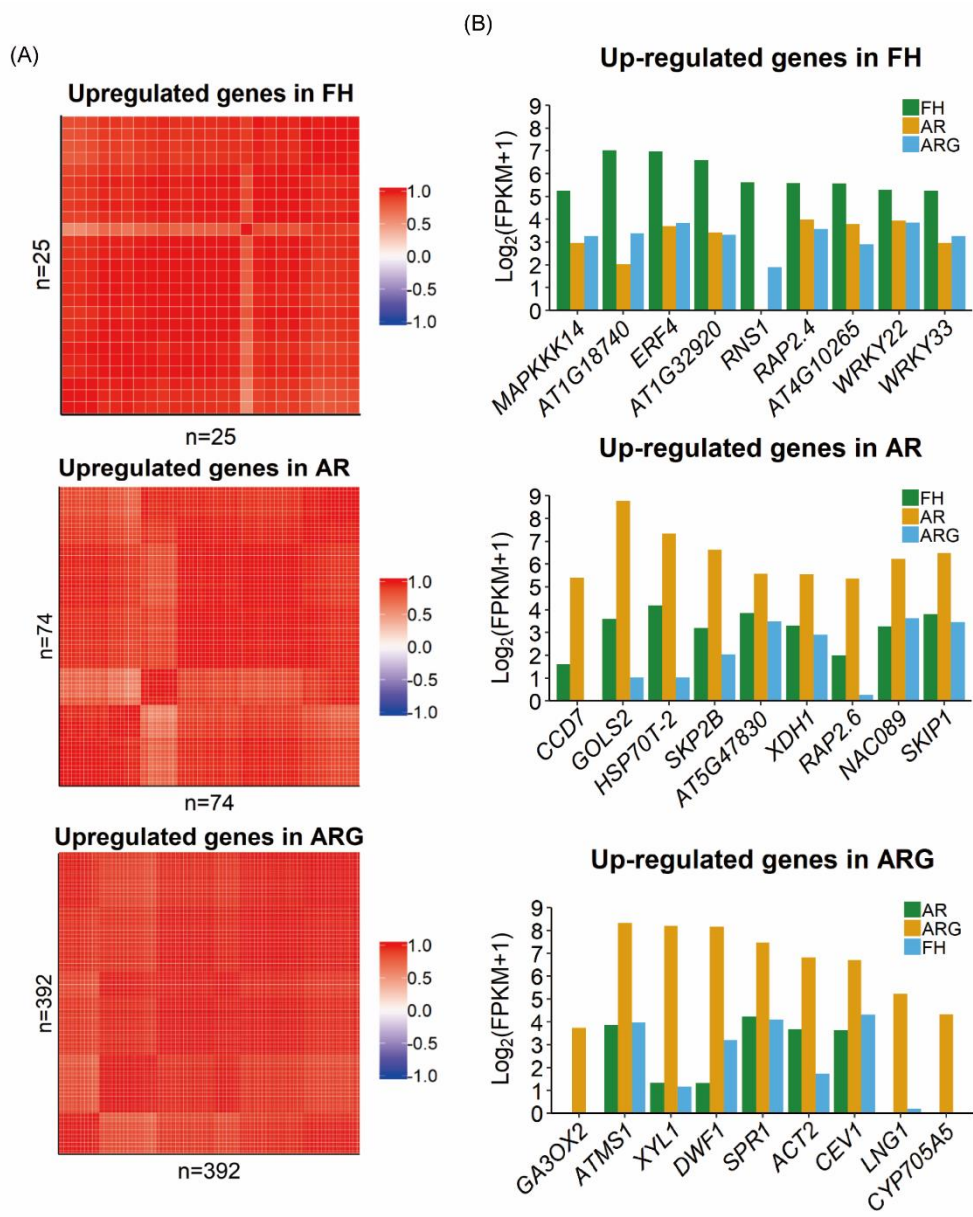
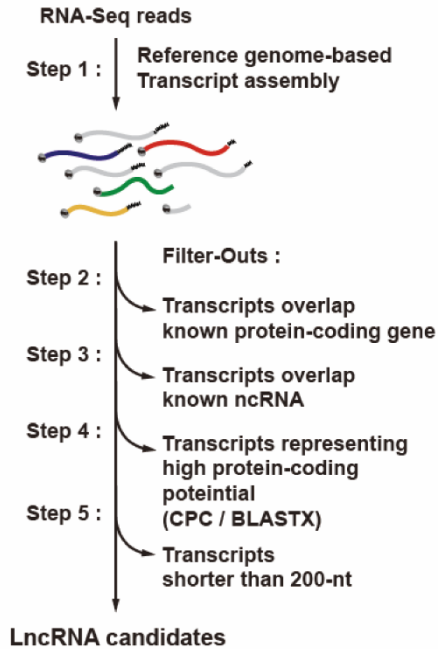


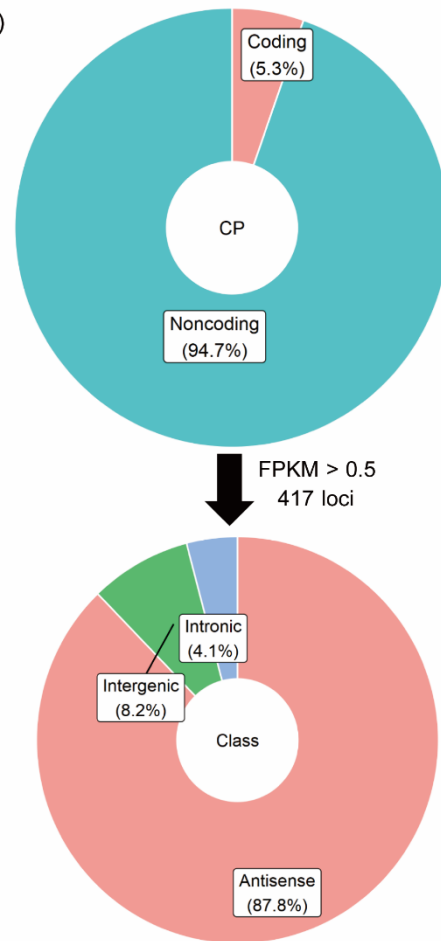
Figure 7. Correlation analysis for novel candidates. (A) Heatmap representation of correlation among annotated genes in developmental stages. (B) Comparison of expression level between marker gene and putative candidates.

(A)

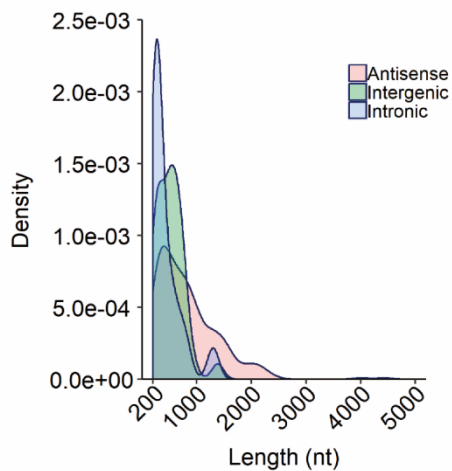
LncRNA identification



(B)



(C)



(D)

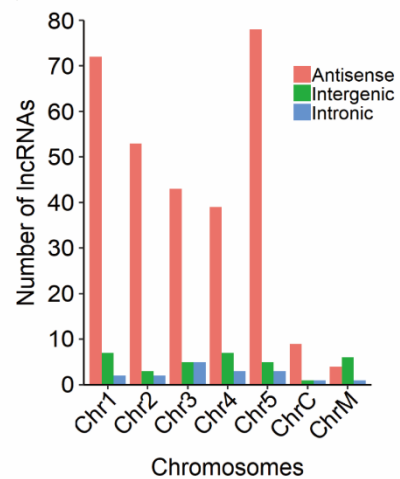


Figure 8. Profiling of lncRNA in *A. thaliana* seeds. The subgroups of transcripts were classified as antisense, intergenic and intronic. (A) scheme for identification of lncRNA (B) Proportion of identified lncRNA. top, the result of coding potential. bottom, proportion of lncRNA (C) Number of lncRNA on *A. thaliana* chromosomes (D) Length distribution of lncRNA.

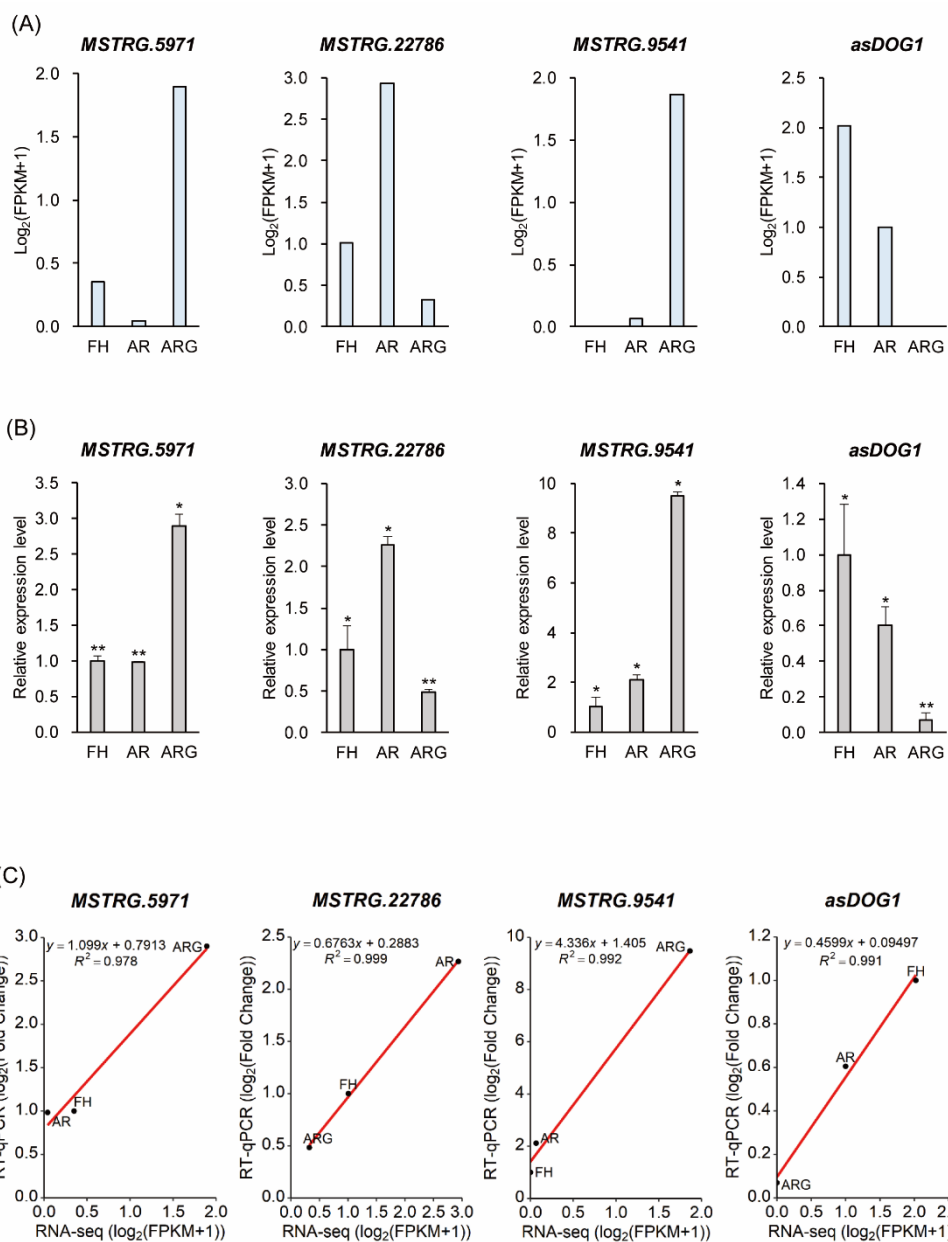


Figure 9. RT-qPCR analysis for validation of representative lncRNAs. (A) $\text{Log}_2(\text{FPKM}+1)$ counts of lncRNAs on RNA-seq. (B) the result of RT-qPCR for validation of representative lncRNAs in developmental stages (C) The result of regression analysis between RNA-seq data and RT-qPCR results. Red line means regression line.

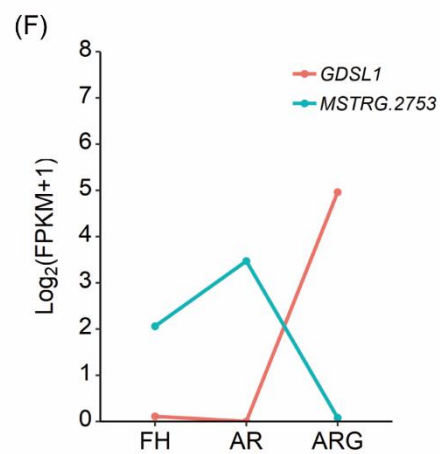
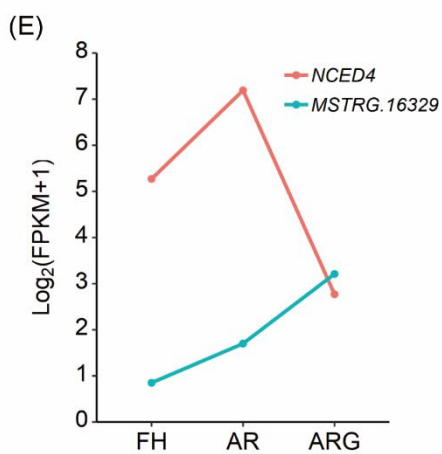
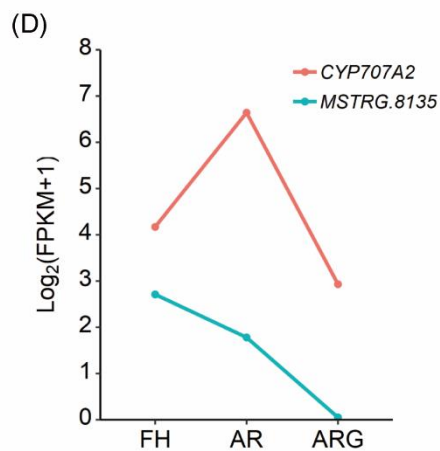
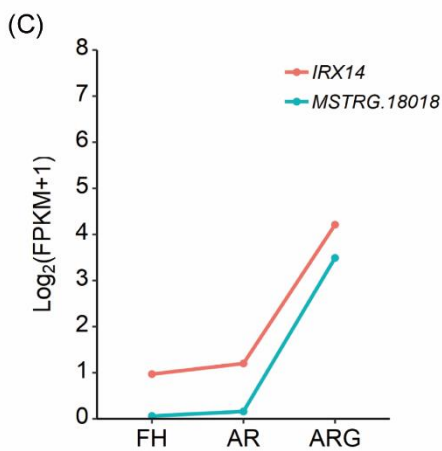
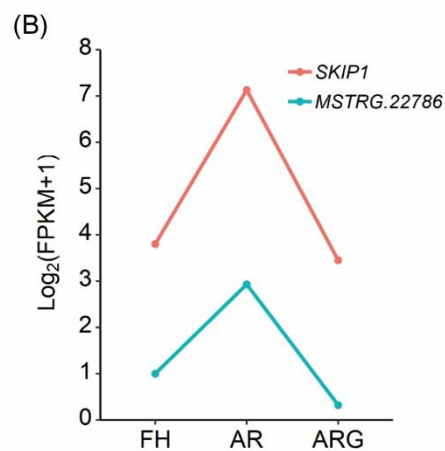
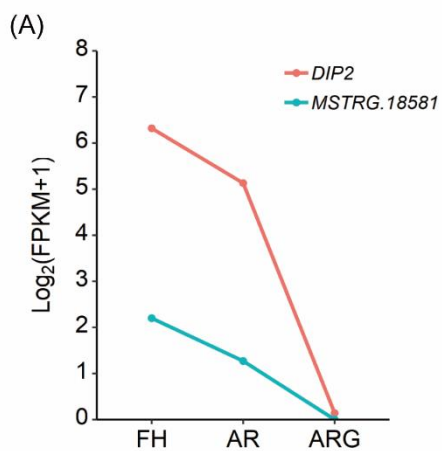


Figure 10. Reciprocal pattern analysis between protein-coding gene and lncRNA.

(A)(B)(C) patterns showing positive correlation and (D)(E)(F) patterns showing negative correlation ($p < 0.01$).

TABLES

Table 1. List of primer sequences used in strand-specific RNA-seq library preparation

Name	Sequence (5' to 3')
TruSeq PCR2.1	CAAGCAGAAGACGGCATACGAGATCGTGATGTG- -ACTGGAGTTCAGACGTGTGCTCTTCCGATC*T
TruSeq PCR2.2	CAAGCAGAAGACGGCATACGAGATACATCGGTG- -ACTGGAGTTCAGACGTGTGCTCTTCCGATC*T
TruSeq PCR2.3	CAAGCAGAAGACGGCATACGAGATGCCTAAGTG- -ACTGGAGTTCAGACGTGTGCTCTTCCGATC*T
TruSeq PCR2.4	CAAGCAGAAGACGGCATACGAGATTGGTCAGTG- -ACTGGAGTTCAGACGTGTGCTCTTCCGATC*T
TruSeq PCR2.5	CAAGCAGAAGACGGCATACGAGATCACTGTGTG- -ACTGGAGTTCAGACGTGTGCTCTTCCGATC*T
TruSeq PCR2.6	CAAGCAGAAGACGGCATACGAGATATTGGCGTG- -ACTGGAGTTCAGACGTGTGCTCTTCCGATC*T
TruSeq_Upper adapter	ACACTCTTTCCCTACACGAC<u>GCTCTTCCGATC</u>*T
TruSeq_Lower adapter	p<u>GATCGGAAGAGCACACGTCTGAACTCCAGTCA</u>*C

(p, monophosphate; *, phosphorothioate; underline, complementary sequence)

Table 2. Read distribution of strand specific RNA-seq library

Library	Raw Reads	Aligned pair	Percentages (%)
Col-0_FH_#1	48,264,130	37,638,660	78.01
Col-0_AR_#1	47,065,716	36,576,037	77.71
Col-0_ARG_#1	71,703,675	53,370,472	74.43

Table 3. List of primer sequences used in RT-qPCR validation

Name	Sequence (5' to 3')
ATqPCR_011	GATTCTTGTGCCGGAGTC
ATqPCR_012	CGTGTCGTAATATGATTTGGTTG
ATqPCR_017	GATGAATCTTTGGTGCAGCTC
ATqPCR_018	GTCTTTACATTGGTCGAGTCC
ATqPCR_029	TTCTATGGCTACGAGTCTGA
ATqPCR_030	CTCGTGACAGTAACTGATATC
ATqPCR_031	CGCACCGTACTGACTACCGA
ATqPCR_032	GGCTCGTTTATGCTTTGTGTGGG

Table 4. List of representative or putative genes in developmental stages

(A)

Symbol	TAIR ID	R	Biological process
AT1G18740	AT1G18740	0.97	ROH1, putative (DUF793)
ERF4	AT3G15210	1	positive or negative ethylene-activated signaling path-way, response to abscisic acid, etc.
AT1G32920	AT1G32920	1	hypothetical protein
RNS1	AT2G02990	1	cellular response to phosphate starvation, response to wounding, RNA catabolic process, etc.
RAP2.4	AT1G78080	0.97	ethylene-activated signaling pathway, response to abiotic stresses, etc.
WIP3	AT4G10265	0.94	response to wounding
WRKY22	AT4G01250	1	defense response, leaf senescence, response to chitin
WRKY33	AT2G38470	1	camalexin biosynthetic process, positive regulation of autophagy, response to abiotic stresses

(B)

Symbol	TAIR ID	R	Biological process
GOLS2	AT1G56600	1	galactose metabolic process, response to abscisic acid, response to water deprivation, etc.
HSP70T-2	AT2G32120	0.97	cellular response to heat, protein refolding, response to heat, etc.
SKP2B	AT1G77000	1	heat acclimation, negative regulation of lateral root development, protein ubiquitination, etc.
AT5G47830	AT5G47830	0.99	hypothetical protein
XDH1	AT4G34890	0.99	purine nucleobase catabolic process, response to ROS, xanthine metabolic process, etc.
RAP2.6	AT1G43160	1	cellular response to heat, ethylene-activated signaling pathway, response to abiotic stresses
NAC089	AT5G22290	0.95	negative regulation of flower development, plant-type hypersensitive response, etc.
SKIP1	AT5G57900	0.96	protein ubiquitination, SCF-dependent proteasomal ubiquitin-dependent protein catabolic process

(C)

Symbol	TAIR ID	R	Biological process
ATMS1	AT5G17920	1	response to cadmium ion, response to salt stress, response to zinc ion
XYL1	AT1G68560	1	cell wall organization, response to cadmium ion, xylan catabolic process
DWF1	AT3G19820	0.97	brassinosteroid biosynthetic process, lignin metabolic process, etc
SPR1	AT2G03680	1	anisotropic cell growth, cellular response to salt stress, etc.
ACT2	AT3G18780	0.92	response to cytokinin, response to far red light, root epidermal cell differentiation
CEV1	AT5G05170	0.98	cellulose biosynthetic process, defense response, cell wall organization, etc.
LNG1	AT5G15580	1	regulation of monopolar cell growth, unidimensional cell growth
CYP705A5	AT5G47990	1	flavonol metabolic process, positive gravitropism, root development

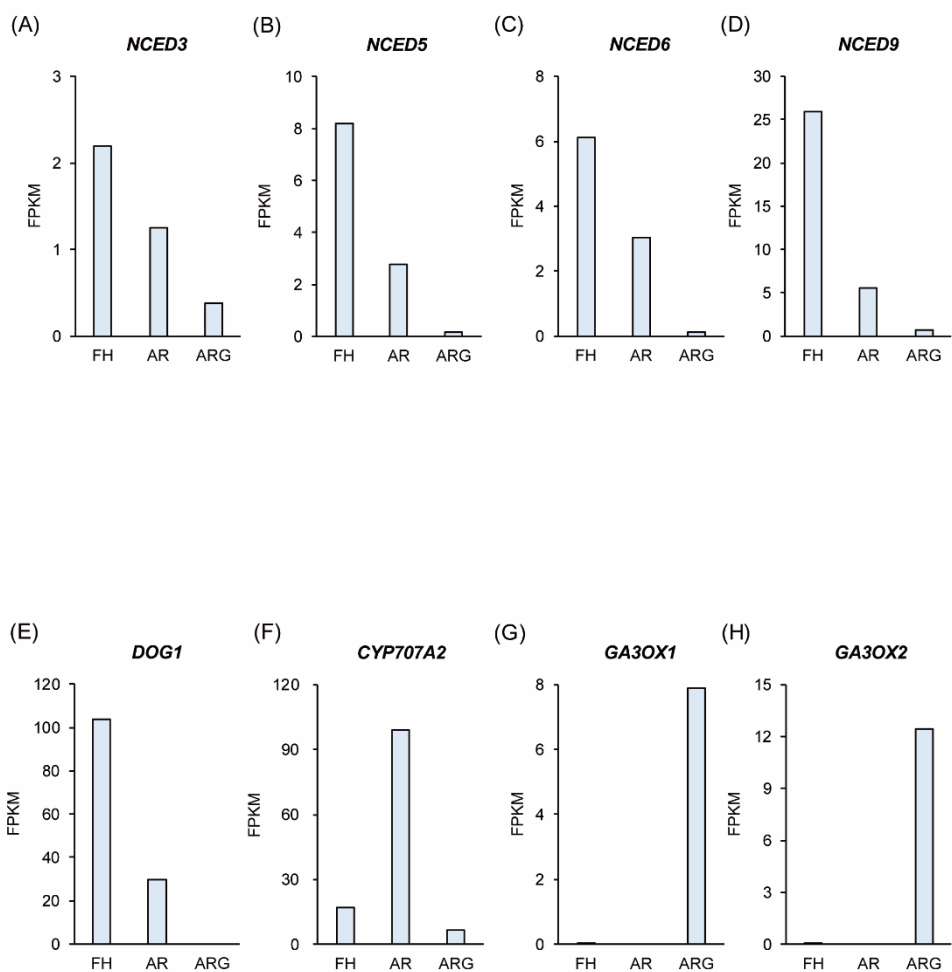
REFERENCE

- Hoffmann MH (2002) Biogeography of *Arabidopsis thaliana* (L.) Heynh. (Brassicaceae). *Journal of Biogeography*. 29: 125–134. doi:10.1046/j.1365-2699.2002.00647.x.
- Mitchell-Olds T (December 2001) *Arabidopsis thaliana* and its wild relatives: a model system for ecology and evolution. *Trends in Ecology & Evolution*. 16 (12): 693–700. doi:10.1016/s0169-5347(01)02291-1.
- Bewley JD. (1994) *Seeds; Physiology of Development and Germination*. New York: Plenum Press, USA.
- Alonso-Blanco, C., Bentsink, L., Hanhart, C. J., Blankestijn-de Vries, H., & Koornneef, M (2003) Analysis of natural allelic variation at seed dormancy loci of *Arabidopsis thaliana*. *Genetics*, 164(2), 711–729.
- Bentsink, L., et al. (2006) Cloning of DOG1, a quantitative trait locus controlling seed dormancy in *Arabidopsis*." *Proceedings of the National Academy of Sciences* 103(45): 17042.
- Bentsink, L., et al. (2010) Natural variation for seed dormancy in *Arabidopsis* is regulated by additive genetic and molecular pathways. 107(9): 4264-4269.
- Graeber, K. et al. (2014) DELAY OF GERMINATION 1 mediates a conserved coat-dormancy mechanism for the temperature- and gibberellin-dependent control of seed germination. *Proceedings of the National Academy of Sciences of the United States of America*, 111(34), E3571–E3580. doi:10.1073/pnas.1403851111
- Morris, K. V. and J. S. Mattick (2014) The rise of regulatory RNA. *Nature Reviews Genetics* 15: 423.
- Hombach, S. and M. Kretz (2016) Non-coding RNAs: Classification, Biology and Functioning. *Non-coding RNAs in Colorectal Cancer*. O. Slaby and G. A. Calin. Cham, Springer International Publishing: 3-17.
- Zhao, Y., et al. (2007) Dysregulation of Cardiogenesis, Cardiac Conduction, and Cell Cycle in Mice Lacking miRNA-1-2. *Cell* 129(2): 303-317.
- van Rooij, E., et al. (2007) Control of Stress-Dependent Cardiac Growth and Gene Expression by a MicroRNA. 316(5824): 575-579.
- Calin, G. A., et al. (2002) Frequent deletions and down-regulation of micro- RNA genes *miR15* and *miR16* at 13q14 in chronic lymphocytic leukemia." 99(24): 15524-15529.
- Wang, W. X., et al. (2008) The expression of microRNA miR-107 decreases early in Alzheimer's disease and may accelerate disease progression through regulation of beta-site amyloid precursor protein-cleaving enzyme 1. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 28(5), 1213–1223. doi:10.1523/JNEUROSCI.5065-07.2008
- Mercer, T. R., et al. (2009). Long non-coding RNAs: insights into functions. *Nature Reviews Genetics* 10: 155.

- Wilusz, J. E., Sunwoo, H., & Spector, D. L (2009) Long noncoding RNAs: functional surprises from the RNA world. *Genes & development*, 23(13), 1494–1504. doi:10.1101/gad.1800909
- Briggs, James A., et al. (2015) Mechanisms of Long Non-coding RNAs in Mammalian Nervous System Development, Plasticity, Disease, and Evolution. *Neuron* 88(5): 861-877.
- Iyer, M. K., et al. (2015) The landscape of long noncoding RNAs in the human transcriptome. *Nature genetics*, 47(3), 199–208. doi:10.1038/ng.3192
- Cheng, J., et al. (2005). Transcriptional Maps of 10 Human Chromosomes at 5-Nucleotide Resolution. 308(5725): 1149-1154.
- Katayama, S., et al. (2005) Antisense Transcription in the Mammalian Transcriptome. 309(5740): 1564-1566.
- Henz, S. R., et al. (2007) Distinct Expression Patterns of Natural Antisense Transcripts in Arabidopsis. 144(3): 1247-1255.
- Wang, Y., et al. (2014) Arabidopsis noncoding RNA mediates control of photomorphogenesis by red light. *Proceedings of the National Academy of Sciences of the United States of America*, 111(28), 10359–10364. doi:10.1073/pnas.1409457111
- Ma, K., Shi, W., Xu, M., Liu, J., & Zhang, F. (2018) Genome-Wide Identification and Characterization of Long Non-Coding RNA in Wheat Roots in Response to Ca²⁺ Channel Blocker. *Frontiers in plant science*, 9, 244. doi:10.3389/fpls.2018.00244
- Hajjari, M., Mowla, S. J., & Faghihi, M. A. (2016). Editorial: Molecular Function and Regulation of Non-coding RNAs in Multifactorial Diseases. *Frontiers in genetics*, 7, 9. doi:10.3389/fgene.2016.00009
- Wang, X.-J., et al. (2005). Genome-wide prediction and identification of cis-natural antisense transcripts in Arabidopsis thaliana. 6(4): R30.
- Wang, H., et al. (2014). Genome-wide identification of long noncoding natural antisense transcripts and their responses to light in Arabidopsis. *Genome research*, 24(3), 444–453. doi:10.1101/gr.165555.113
- Li, L., et al. (2014). Genome-wide discovery and characterization of maize long non-coding RNAs. 15(2): R40.
- Wang, H., et al. (2015). Analysis of non-coding transcriptome in rice and maize uncovers roles of conserved lncRNAs associated with agriculture traits. 84(2): 404-416.
- Lu, T., et al. (2012). Strand-specific RNA-seq reveals widespread occurrence of novel cis-natural antisense transcripts in rice. 13(1): 721.
- Heo, J. B. and S. Sung (2011) Vernalization-Mediated Epigenetic Silencing by a Long Intronic Noncoding RNA. 331(6013): 76-79.
- Kim, D.-H. and S. Sung (2017) Vernalization-Triggered Intragenic Chromatin Loop Formation by Long Noncoding RNAs. *Developmental Cell* 40(3): 302-312.e304.
- Qin, T., Zhao, H., Cui, P., Albeshier, N., & Xiong, L. (2017). A Nucleus-Localized Long Non-Coding RNA Enhances Drought and Salt Stress Tolerance. *Plant physiology*, 175(3), 1321–1336. doi:10.1104/pp.17.00574

- Zhong, S., et al. (2011) High-Throughput Illumina Strand-Specific RNA Sequencing Library Preparation." 2011(8): pdb.prot5652.
- Lamesch, P., et al. (2011) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools." Nucleic Acids Research 40(D1): D1202-D1210.
- Pertea, M., et al. (2016) Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature Protocols 11: 1650.
- Pertea, M., et al. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads." Nature Biotechnology 33: 290.
- Di, C., et al. (2014) Characterization of stress-responsive lncRNAs in Arabidopsis thaliana by integrating expression, epigenetic and structural features. 80(5): 848-861.
- Cheng, C.-Y., et al. (2017) Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. 89(4): 789-804.
- Kalvari, I., et al. (2017) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Research 46(D1): D335-D342.
- Nawrocki, E. P. and S. R. Eddy (2013) Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29(22): 2933-2935.
- Kang, Y.-J., et al. (2017). CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Research 45(W1): W12-W16.
- Storey, J. D. (2002). A direct approach to false discovery rates. 64(3): 479-498.
- Yu, G., Wang, L. G., Han, Y., & He, Q. Y. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. Omics: a journal of integrative biology, 16(5), 284–287. doi:10.1089/omi.2011.0118
- El Hadrami, A., Adam, L. R., El Hadrami, I., & Daayf, F. (2010) Chitosan in plant protection. Marine drugs, 8(4), 968–987. doi:10.3390/md8040968
- Narayana Murthy, U. M. and W. Q. Sun (2000) Protein modification by Amadori and Maillard reactions during seed storage: roles of sugar hydrolysis and lipid peroxidation. Journal of Experimental Botany 51(348): 1221-1228.
- Fedak, H., et al. (2016) Control of seed dormancy in Arabidopsis by a cis-acting noncoding antisense transcript. 113(48): E7846-E7855.

SUPPLEMENTARY FIGURES AND LEGENDS



Supplementary figure 1. Expression of marker gene in *Arabidopsis thaliana* seeds on RNA-seq. FPKM value of (A)(B)(C)(D) *NCED* family genes, (E) *DOG1*, (F) *CYP707A2* and (G)(H) *GA3OX* family genes.

ABSTRACT IN KOREAN

애기 장대 종자 발달에 따른 유전체 프로파일링

종자의 발달은 역사적으로 생존과 번식을 위해 끊임없이 생태환경에 따라 휴면과 발아를 거듭하며 생존해왔다. 이는 하나의 온전한 식물로 성장하기 위한 필수적인 현상이다. 종자 발달의 현상을 결정하는 중요한 요소로는 크게 환경과 이에 반응하는 호르몬으로 구성된다. 그러나 휴면과 발아는 하나의 독립적인 메커니즘에 의한 현상으로 나타나기 보다 유전자 간 상호 관계에 의해 발생하는 현상이다. 이를 규명하기 위해 많은 연구진들이 애기장대에서 종자 발달 관련 연구의 필요성 및 사례가 증가하는 추세이다.

최근에는 miRNA 및 lncRNA 등과 같이 단백질로 암호화되지 않는 non-coding RNA 에 대한 다양한 연구를 통해 표적 유전자의 상위 조절 인자로서 동 식물 내 중요한 현상을 발생시키는 다양한 메커니즘에 관여하는 사례가 보고되고 있다. early stage 에서 cold exposure 에 따른 애기장대에서는 FLOWERING LOCUS C (FLC) locus 에서 전사되는 COLDAIR 가 보고되었으며, drought 및 salt 스트레스를 포함한 다양한 스트레스 환경에 대한 반응 및 이에 대한 대응 기작에 관여하는 것으로 알려져 있다. 또한 최근에는 DOG1 에 대한 asDOG1 이 밝혀져 종자의 휴면과 발달에 대한 lncRNA 연구에 많은 진전을 보이고 있다.

본 연구는 종자 발달 단계에 따른 전사체 수준의 변화에 대한 유전자 간 발현 정도를 관찰하고자 strand-specific RNA-seq 기법이 사용하여 유전자 발현 분석 및 전사체를 분석하였다. 또한 putative long-non coding RNA deriving loci 발굴을 수행하였다. 이를 통해 휴면 및 발아를 확인하기 위한 마커 유전자 이외의 새로운 마커 유전자로 사용될 가능성 있는 유전자에 대한 다양한 분석을 실행하였다. 위의 연구를 통해

확인된 표지 유전자는 애기장대 이외에 모든 식물에 걸쳐 종자 발달 단계에 따른 전사체 수준의 변화에 대한 이해를 도울 수 있음을 기대한다. 또한 protein-coding gene 과 long non-coding RNA 간 세포 내 역할을 규명하는 연구를 진행하는 데에 있어 기초 자료로 활용될 수 있음을 기대한다.

